Journal of Education and Learning; Vol. 2, No. 2; 2013
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

# Merging Large-Scale Assessment Data for Secondary Analysis: Experiences with EQAO's Data

Gul Shahzad Sarwar[1], Carlos Zerpa[2], Christina van Barneveld[2], Marielle Simon[1] & Karieann Brinson[2]

[1] Faculty of Education, University of Ottawa, Ottawa, ON, Canada

[2] Faculty of Education, University of Lakehead, Thunder Bay, ON, Canada

Correspondence: Gul Shahzad Sarwar, Faculty of Education, University of Ottawa, Ottawa, ON, K1N 6N5, Canada. Tel: 1-613-562-5804. E-mail: gsarw024@uottawa.ca

Received: January 25, 2013 Accepted: February 22, 2013 Online Published: April 2, 2013

doi:10.5539/jel.v2n2p44 URL: http://dx.doi.org/10.5539/jel.v2n2p44

*The research was financed by the Social Sciences and Humanities Research Council of Canada.*

**Abstract**

This paper is a narrative of our experience in analyzing and merging data files provided to us by the Education Quality and Accountability Office (EQAO). In the paper, we propose a scheme of merging data files by means of Structured Query Language (SQL, pronounced as "sequel"). Although, the narrative of our experiences using this merging scheme could have been extended to any number of data files, the aim of this work was to merge only three EQAO data files. Via this merge process, we were able to gain meaningful information and facilitate the analysis of EQAO data to answer our research questions. By using SQL queries, our approach was not only to analyze the available data files but also to construct a narrative about viewing and handling data contained in the files.

**Keywords:** data merging, secondary data analysis, large-scale assessment, EQAO, narrative

## 1. Introduction

In a traditional model of quantitative data analysis, it is expected that a researcher considers data file sources separately (Voronin, 2006). This approach may be beneficial in obtaining higher degrees of freedom from the data, because when data files are merged, some records in the files are generally lost. While the traditional model of data analysis has an advantage in its simplicity, it does not address the case when some helpful explanatory information about a variable is present in different data files. This information could only become meaningful if the data files were merged.

Data merging is a procedure for integrating different sets of data from multiple files into one, on the basis of one or more key variables. The two main advantages of merging data file sources are (a) an increase in the number of variables which leads to a gain of related information and (b) the possibility of obtaining new results, which were not initially planned prior to data collection. The increased number of variable size may sometimes result in a reduced data sample size which, in turn, decreases the power of statistical tests for observing the effects. On the other hand, an increased number of variables may raise the possibilities of answering new research questions or triangulating some of the results. While merging data file sources raises the possibilities of getting meaningful results during data analysis, the merging process itself may be problematic. For instance, there is a large potential for decreasing the quality of the merged data, even below the level of the original sources (Naumann & Häussle, 2002). The decrease in quality of the merged data is mainly due to data conflicts in the source files. Data sources enforce each other when they store the same data for an attribute about a particular object. For example, two data sources reporting that the same student identification number (StudentID) obtained the same marks enforces that this student has scored the said marks. Data sources complement each other if they store data about different attributes of the same object. For example, a data source storing only StudentID and Course_Code complements another data source, which stores StudentID, Name and Year of a student. Data sources conflict if they store different data for an attribute about a particular object. For example, a data source stores a mark of 97 for the StudentID 501 conflicts with the data source that stores a mark of 95 for the same StudentID. Resolving such conflict is challenging for a researcher, and generally, this type of data is excluded from the merged file.

The process of merging data file sources from two data files having one-to-one relation may be accomplished in a straightforward manner by simply sorting the data files, merging them and eliminating duplicate records (Bitton & DeWitt, 1983). The process becomes more complicated, however, when data files involve one-to-many or many-to-many relationships, or when entities are represented differently in available data files. The problem of merging data files containing heterogeneous information is considered a serious concern for many research organizations. In the literature, instances of this problem are also called *record linkage* (Fellegi & Sunter, 1969), *semantic integration problem* (ACM, 1991) or *instance identification problem* (Wang & Madnick, 1989). Business organizations call this problem the *merge/purge problem* (Hernández & Stolfo, 1998). The issue is normally encountered in education research when a researcher conducts secondary analysis of data originally collected for a multi-method or multilevel study. With a growing body of research on large-scale assessment and large-scale surveys, the problem is becoming increasingly important for researchers analyzing large amount of data from such multi-method or multilevel studies. The process of merging data file sources involves many steps. The process begins from the initial database integration to data cleaning and finally ends at the actual data merging. While most of the research is concentrated on the initial steps of data file merging, there are only a few studies about actual and practical merging of the data file sources using query against multiple data files (Naumann & Häussle, 2002). Further, these studies do not fully describe in detail the process of merging and verification of data file sources.

This paper presents a rich account of challenges faced by our research team during the process of merging of data from the Education Quality and Accountability Office (EQAO), in Ontario, Canada for secondary analysis to answer some specific research questions related to an assessment of Grade 9 mathematics. The paper focuses on Microsoft SQL Server 2008 while writing SQL queries to merge data file sources and to check data integrity of the merged files. Occasionally, the research group members also used SPSS 19 to merge data files because it is easily accessible in an educational environment and relatively user-friendly. Using the syntax function of SPSS which can automatically create a re-executable syntax file, researchers can easily save the commands without writing a script and modifying the file later on. But for many advanced functions, the SQL is preferable to SPSS, specifically when data files involve one-to-many relationships. For example in our case, the data provided to us contained one-to many relationships between a teacher and his students. In other situations, SQL is preferable when entities are represented differently in available data files. For example, in the data files provided to our research group, the variable "Program" in the student questionnaire was considered as numeric data type, whereas the equivalent variable in the teacher questionnaire was named "AppliedOrAcademic" and was alphanumeric. MySQL is another commonly used open source database management software, which is freeware and cross platform. All SQL queries written by our research group using Microsoft SQL Server 2008 can also be used as is in MySQL and will produce similar results with the similar data. We preferred Microsoft SQL Server 2008, however, because it is reliable, versatile, and high performing in terms of its capabilities in development and management of a database (Thakar, Szalav, Fekete, & Gray, 2008).

## 2. Secondary Analysis of Educational Data

Secondary use of educational data has a vital role in improving knowledge about the educational system and learning process of a student. Secondary analysis of data is classified as a method in which a researcher uses existing data to answer research questions that may or may not have been proposed when the data were originally collected (Rew, Koniak-Griffin, Lewis, Miles, & O'Sullivan, 2000). The secondary data analysis may be done by the original researcher or agency who collected the data, or it may be done by other researchers (Herron, 1989). Given the large amount of data collected, much of these data remain under-analyzed (Crocker, 2002). As a consequence, important issues may remain unexplored or partially explored. For example, Nagy, Demeris and van Barneveld (2000) argued that educational indicators data or contextual data collected from multiple stakeholders (e.g., teachers questionnaires, students questionnaires, principal questionnaires, parent questionnaires) as a part of large-scale assessment programs remain under-used, when making interpretations about student learning and achievement results. They further proposed to examine the relationships between achievement data and contextual or background data to better understand the learning process of a student.

In literature, a variety of advantages are associated with secondary analysis of data (Simon, Roberts, Tierney, & Forgette-Giroux, 2007). An investigation of existing data allows us to identify new research questions. New research questions can be formulated from existing data sets or by merging two or more data sets. For example, Wolfe, Ray and Harris (2004) conducted secondary analysis of data from The National Center for Educational Statistics, which is responsible for large-scale surveys of educational institutions in the United States. They combined information across several items within the various questionnaires used in the national surveys to discuss three types of teacher perception: perception of influence, perception of students, and school climate. They further mentioned that it is important to verify that measures used to combine the data exhibit sufficient quality to warrant

their use during the process of data analysis. Research designs featuring secondary analysis are also cost-effective and less time consuming than those requiring actual data collection (Rew et al., 2000). Data collection is particularly expensive when a researcher has to gather data from multiple sources. Application of secondary analysis enables a researcher to devote more resources toward other stages of the research process. Although access to some types of information may sometimes require a fee, many of the data sets the researchers need are available free of charge (Gorard, 2002). Another advantage of secondary data analysis is that the sample size is generally very large when data are collected from demographically diverse population (Hofferth, 2005). In such a situation, researchers have the potential for studying larger and more representative samples of the population, which increases the generalizability of findings and statistical power (Moriarty et al., 1999). These large samples also allow the use of more sophisticated statistical methods and give a researcher the ability to study subsamples of interest. For instance, a researcher can compare attitudes toward homework of a subsample of students in French language schools with those in English language schools within the same educational jurisdiction.

Apart from several advantages, there are practical and methodological issues related to maintaining consistency between new research questions and the original data when conducting secondary analysis (Coyer & Gallo, 2005). One such requirement of consistency is that the definition of all variables included in the new research questions should be the same as in the original data. For instance, the definition of the motivation of a student in the original data should be consistent with the theoretical framework of the secondary analysis of data. The unit of analysis such as a student, teacher and classroom or school also needs to be same in both studies (Moriarty et al., 1999). Finally, the researcher needs to assess the accuracy, completeness, and missing data in the sample.

## 3. Large-Scale Assessment and the EQAO Project

Large-scale assessments are an important part of educational accountability systems (Lane & Stone, 2002). In the province of Ontario, Canada, the government ensures greater accountability in the educational system via the EQAO, which is an agency of the Ontario Ministry of Education whose mission is to enhance the quality of education in Ontario. To accomplish this task, the EQAO conducts yearly a large-scale assessment of all Grade nine students in mathematics. Given that there are semestered and non-semestered schools, private and public schools, English-language and French-language schools, and two different programs (applied and academic), different versions of the mathematics test are developed and administered separately. Large-scale assessments also include background questionnaires for students, teachers and school principals in order to examine the various relationships that may exist among their respective perspectives and student achievement on the tests. These various factors lead to the construction of separate data sets, each with their own sets of variables. One of the main variables of interest in this study was the fact that teachers in the schools administer this provincial level assessment and mark some of its components for use in the final grade of students in mathematics courses. The selection of EQAO test items to be assessed and their relative weight, which can vary from 0% to 30% towards students' grades, is a teacher, school or board decision. We were particularly interested in looking at the relationship between this practice and student achievement on the EQAO mathematics test. In October 2008, the EQAO Research Committee expressed interest in our research and provided us with access to anonymous data sets of the 2010 Grade nine assessment of mathematics. These data sets contained EQAO results of the entire population of Grade nine students in Ontario. Apart from some qualitative data and some files containing the definitions of variables, three main quantitative data files were:

1) G9_2010_ItemResponses.csv, comprised of student responses to the mathematics test items,
2) G9_2010_ISD_SQd.csv, comprised of student responses to the student questionnaire items, and
3) G9_2010_TQ.csv, containing teacher responses to the teacher questionnaire items.

The quantitative data files provided by EQAO had some overlap of key variables, which allowed us to merge these files and to subsequently conduct secondary data analysis. The targeted variables were scattered over the three files, and it was essential for us to link data from the three data files simultaneously for the analysis. Therefore, we were required to model data collected using different instruments and respondents (both students and teachers) into an integrated file. The three main objectives while merging data files were to:

1) Merge student data from G9_2010_ItemResponses.csv with student data from G9_2010_ISD_SQd.csv by RecID;

2) Merge teacher data from G9_2010_TQ.csv with the output of the previously merged student data by ClassID; and

3) Merge teacher data from G9_2010_TQ.csv with the output of the previously merged student data by ClassID and Program/AppliedOrAcademic.

## 4. Analysis of Individual Data Files Using SQL Queries

To address concerns and research questions based on the outcome of the large-scale assessment, we were initially faced with the challenges of accurately merging the data. To begin with, we needed to select efficient and proper merging techniques (one-to-one relation, one-to-many relations, many-to-many relations) to combine the EQAO data file sources from students' self-report data, students' item response data and teachers' self-report data. Since there is an assumption that the outcomes of the large-scale assessment conducted by EQAO are an accurate reflection of each student's knowledge and skills, the use of proper data merging techniques will facilitate the necessary information required to address research questions that are geared toward studying the impact on students' motivation when a portion of the large-scale assessment may count or may not count toward their academic grades.

Proper data merging techniques are also important when combining large-scale assessment data sources to ensure consistency among relations, to condition the data and to filter invalid cases that may skew the results. In addition, we needed to bear in mind that students, parents, teachers, school administrators, government officials and other stakeholders interpret data from large-scale assessments to plan for improvement, guide decision-making, allocate resources and/or set educational policies (Coburn & Talbert, 2006; Kornhaber, 2004; van Barneveld, Stienstra & Stewart, 2006). Thus, the strategy of having some or all of the items on a large-scale assessment count towards course grades needed to be examined thoroughly to determine its overall effectiveness by first properly merging the data.

Our research group began by writing SPSS 19 scripts to sort, filter and merge data files. When the data files needed to merge are large (i.e., containing hundred thousands of records), the process of merging in SPSS may become time consuming and difficult to handle because it requires the user to perform more steps or develop large SPSS scripts to condition and merge the data. Therefore, it is appropriate to select the variables of interest (using "drop" or "keep" commands) and then proceed to merge data files. The output file obtained after the merging process contains variables of a primary file plus scores of variables from the secondary file that could be used for measuring the required effects. In SPSS terms, this process is an "add variables" merge process. For the merge algorithm to be correctly executed in SPSS, one must make sure that all the records in a data file are sorted by a primary key. Our group thus joined student data files G9_2010_ItemResponses.csv and G9_2010_ISD_SQd.csv by RecID as a primary key using SPSS. RecID was a numeric data type. Therefore, we did not face any difficulty in merging the two files as most of the merging techniques are very efficient when dealing with numeric values. Alphabetic string data, for example, name of a participant or address of a participant, in a common field of tables present additional challenges because these items might be different among different data tables because of typos or capitalization. Hence, the equality of values over a common "join" attribute may not be specified as a simple arithmetic value. The process of merging files using SPSS became more difficult for us when the task was to join three files because it requires the user to analyze the data and apply manual filtering to drop invalid cases in the process of merging data files. We concluded that Microsoft SQL Server 2008 was a more efficient way to handle the task. We wanted to analyze and merge data files as efficiently as possible with a minimum amount of coding. Understanding how data are organized is a key element in the research process, but better technical choices can help to avoid unnecessary technical efforts and to obtain comprehensive results. The key element in the process of merging files is the knowledge of how different variables are related to each other, what data type they have, and does the same variable exist in another data file. Therefore, using SQL queries, our group first analyzed the three individual files provided by EQAO. This analysis provided us an overview of data contained in the files including the number of records and the number of variables. A number of queries were written to analyze the three data files provided by EQAO. According to our analysis, the student item responses file G9_2010_ItemResponses.csv had a total 150,186 records (i.e., number of rows). In this file, students were divided into two groups on the basis of the program in which they were enrolled at school—"applied" or "academic". There were 46,002 applied students and 104,184 academic students. We used following SQL query to determine this:

```
SELECT Program, COUNT(*)
FROM G9_2010_ItemResponses
GROUP BY Program
```

Each student in the table was given a unique RecID assigned by EQAO. In our group discussions, one researcher argued that all the RecIDs may not be unique. There may be a data entry error, typo, or duplication of RecIDs. This was an important issue for us as we were considering RecID as a primary key to merge student data sets and for that unique RecID was required. This was again checked by the following SQL query and we found that all

RecIDs were unique:

```
SELECT RecID, COUNT(*)
FROM G9_2010_ItemResponses
GROUP BY RecID
HAVING COUNT(*) > 1
```

The result of this query was empty, which means that RecID in the student item responses file was unique across the table. Contrary to RecID, when we checked uniqueness of StudentID, we got 1,257 duplicate StudentIDs. We used the following SQL query to determine this:

```
SELECT StudentID, COUNT(*)
FROM G9_2010_ItemResponses
GROUP BY StudentID
HAVING COUNT(*) > 1
ORDER BY COUNT(*) DESC
```

There were seven students with the StudentID of "0" and 625 other StudentIDs appeared twice (according to EQAO field definition, StudentID of "0" indicates that a student's identifier is unknown). Although some of these StudentIDs were there because of an error in the data entry, these 1,257 students were still valid cases for us because we were in fact not using StudentID but RecID as the primary key for the student item responses file. If required, these duplicate records can be eliminated from a file using the "traditional" method where the file is first sorted in order to bring all the duplicate records together. Then a sequential pass is made through the file to compare adjacent records and to eliminate all duplicated records. Since most of the traditional database management systems provide a sort facility, this approach is clearly the simplest (Bitton & DeWitt, 1983). In a recent commonly adopted approach, however, duplicates are eliminated as part of query written for merging data files (Hernández & Stolfo, 1998). This approach significantly reduced the execution time to eliminate duplicate records in the output data file.

The student questionnaire file G9_2010_ISD_SQd.csv had a total of 154,569 records. All the RecIDs in the file were also unique. In this data file, students were also divided into two groups on the basis of their program—"applied" or "academic". There were a total of 49,059 students in applied program and 105,510 students in academic program. We used the following SQL query to determine this:

```
SELECT Program, COUNT(*)
FROM G9_2010_ISD_SQd
GROUP BY Program
```

The student questionnaire file was checked for the uniqueness of RecIDs, and we found that all the RecIDs were unique. Again when we check this file for the uniqueness of StudentIDs, we found 1,766 duplicates. There were seven students with the StudentID of "0", one StudentID appeared three times and 878 other StudentIDs appeared twice. We also checked that all the student item responses file's 150,186 records were also in the student questionnaire file, but the later file had 4,383 extra records. This was checked by writing the following SQL query:

```
SELECT *
FROM G9_2010_ItemResponses A
WHERE NOT EXISTS (SELECT * FROM G9_2010_ISD_SQd B WHERE A.RecID = B.RecID)
```

The result of this query was empty, which means that all the student item responses file's records existed in the student questionnaire file.

The teacher questionnaire file G9_2010_TQ.csv had total 6,373 records. In this file, TQUID was a primary key as it was unique for all the records. This teacher data file and previously described student data files had two common columns named "ClassID" and "SchoolID". This enabled us to merge teacher data in this file with student data in the previous two files as one of our research questions' requirement was to analyze student data corresponding to the teacher data. There were 2,087 teachers who filled out the questionnaire for the applied program, and 3,021 who filled it out for the academic program, whereas there were 1,265 teachers who filled out the questionnaire but did not mention their program. In the table given by EQAO, their program was represented

by the symbol "#", means that they did not mention their program in the questionnaire. We used the following SQL query to determine the number of teachers in each program, including the teachers who did not mention their program:

```
SELECT AppliedOrAcademic, COUNT(*)
FROM G9_2010_TQ
GROUP BY AppliedOrAcademic
```

**5. Data Merging and Verifications Using SQL Queries**

Sometimes, combined information from various data sets is required so that educational leaders can make evidence-based decisions. Data file merging is actively applied in such a scenario to extract a maximum amount of useful information from available data sets (Voronin, 2006). Because of the diversity of data in files collected from a large number of schools, traditional forms of data analysis may not generate a result that reflects the holistic knowledge about schools (Louis, 1982). Rather than analyzing each data file separately or merging all the data files together at once, our research group decided to combine only those data files at a time that contained key variables relevant to the research question of interest.

As a *first step,* we joined student item responses file G9_2010_ItemResponses.csv and student questionnaire file G9_2010_ISD_SQd.csv by RecID. The joining was done by writing the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
```

The purpose of the join-approach is to create a resultant file that merges two files on the basis of certain criterion. There were 150,186 records in the student item responses file and 154,569 records in the student questionnaire file. This shows that there were 4,383 extra records in the student questionnaire file as compared to the student item responses file. For example, RecID 1410V167 was in the student questionnaire file but missing in the student item responses file. As stated earlier, we checked that all the 150,186 records in the student item responses file were also found in the student questionnaire file. This means that 4,383 students (3%) filled out the questionnaire but did not write item responses. When we merged the two files, these 4,383 records were excluded from the output file because they did not have related data in the student item responses file. This kind of join is called an "inner join" where one gets only those rows of a table that are common to both the tables. It is the most common join process and is considered as a default join process. Inner join compares each row of a table with each row of another table to find all of the pairs of rows and creates a resultant table where the join condition satisfies (Naumann & Häussle, 2002).

The output file was further checked for the similarity of "ClassID", "SchoolID" and "Program" linked with the "RecID" across a record for the verification of a correct merging output. It was found that they were the same. It was checked by writing the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
WHERE A.ClassID<>B.ClassID or A.SchoolID<>B.SchoolID or A.Program<>B.Program
```

The result of this query was empty, which means that all the "ClassID", "SchoolID" and "Program" data, linked with the "RecID" across all records, were the same.

In the *second step*, the teacher questionnaire file G9_2010_TQ.csv was merged with the combined student data file generated from the merge process at the first step by ClassID. This merge was done by using the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
JOIN G9_2010_TQ C ON A.ClassID = C.ClassID AND B.ClassID = C.ClassID
```

A reliable merged file requires that no data in the source files be dropped during the merging without any predefined criterion. A researcher should know the reason why some of the data are lost. When some data are

lost, the researcher should recheck the reliability of the predefined criterion. In our case, there were 150,186 records in the combined student data file generated at the first step, and after this merge, we had 132,772 records. This shows 17,414 records (12%) were dropped at the second step, probably because these students did not have their corresponding ClassIDs in the teacher questionnaire file. We checked this by writing the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
WHERE NOT EXISTS (SELECT * FROM G9_2010_TQ C WHERE A.ClassID = C.ClassID AND B.ClassID = C.ClassID)
```

Missing data due to item nonresponse in a questionnaire is a common problem when dealing with large-scale studies (Peugh & Enders, 2004; Rubin, 1996). Missing data are observations that are intended to be made but are not gathered for a variety of reasons (Sterne et al., 2009). Rubin (1976) identified three missing data mechanisms. Missing data can be characterized as "missing completely at random" when a missing value on a variable is independent of its own value or of any other value in the database. Missing data can be "missing at random" when a missing value on a variable is independent of its own value but is related to another variable in the database. Finally, missing data may be "missing not at random" when a missing value on a variable is dependent on its own value (Allison, 2001). There are a number of approaches in treating missing data, including deletions, substitutions, imputations, and data modeling. The listwise deletion procedure (in which all data records are excluded where any variable value is missing) is a feasible option and produces unbiased estimates when data items are missing completely at random, when the sample size is large and when the number of missing values is small (Basilevsky, Sabourin, Hum, & Anderson, 1985; Roth & Switzer, 1995; Witta, 1992). In the merged file at the second step, there were 25,175 students' records that had "#" in the column "AppliedOrAcademic". This shows that for the corresponding ClassIDs, the teachers did not mention their program in the teacher questionnaire file. Since these data were "missing completely at random", listwise deletion of these records was the most appropriate approach. These 25,175 students were excluded at the third step as a part of query by adding the condition of the same "Program" in the SQL query.

In the *third step*, the teacher questionnaire file G9_2010_TQ.csv was merged with the combined student data file generated from the merge process at the first step by ClassID, and two equivalent variables Program/ Applied Or Academic. This merge was done by writing the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
JOIN G9_2010_TQ C ON A.ClassID = C.ClassID AND A.Program = C.AppliedOrAcademic
```

There were 132,772 records in the merged data file generated at the second step, and, after this merge, we had 106,124 records. This shows 26,648 records (20%) were dropped at the third step. Altogether, we lost 44,763 records in steps two and three, which represents 29% of the original student questionnaire file.

As stated earlier, there were 25,175 students who had "#" in the column of "AppliedOrAcademic" in the output file at the second step. These 25,175 students were dropped from the merged file at the third step, because combined student data file generated at the first step was merged with the teacher questionnaire file on the criteria of same ClassID and Program/AppliedOrAcademic. We rechecked the number of such students by writing the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
JOIN G9_2010_TQ C ON A.ClassID = C.ClassID AND B.ClassID = C.ClassID
WHERE AppliedOrAcademic = '#'
```

There were 1,473 students' records in the combined student data file generated from the merge process at the first step in which the variable "Program" did not match with the variable "AppliedOrAcademic" in the teacher questionnaire, but their ClassIDs were the same. This might be because of a data entry error. We counted the

number of these students by writing the following SQL query:

```
SELECT *
FROM G9_2010_ISD_SQd A
JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
JOIN G9_2010_TQ C ON A.ClassID = C.ClassID AND B.ClassID = C.ClassID
WHERE AppliedOrAcademic IS NOT NULL AND A.Program<>C.AppliedOrAcademic
```

There were also 193 teachers which were not included in the output merged file at the third step. Their combination of "ClassID" and "AppliedOrAcademic" was different from any of the student's combination of "ClassID" and "Program". This also might be because of a data entry error. We determined the number of missing teachers in the merged file by writing the following SQL query:

```
SELECT *
FROM G9_2010_TQ A
WHERE AppliedOrAcademic IN (1,2)
AND NOT EXISTS
(SELECT TQUID FROM
    (SELECT DISTINCT TQUID
    FROM G9_2010_ISD_SQd A
    JOIN G9_2010_ItemResponses B ON A.RecID = B.RecID
JOIN G9_2010_TQ C ON A.ClassID = C.ClassID AND A.Program = C.AppliedOrAcademic) B WHERE A.TQUID = B.TQUID
```

The results of potential data entry error in the three files provided by EQAO are summarized in Table 1.

Table 1. Results of data integrity of the three data files based on possible data entry errors

| File | Total number of records | Data entry error about a variable [Number of variables] | Number of records having potential data entry error | Percentage of potential data entry error about an item |
|---|---|---|---|---|
| G9_2010_ItemResponses.csv | 150,186 | Duplicate StudentID [1] | 1,257 | 0.84% |
| G9_2010_ItemResponses.csv | 150,186 | Variable "Program" does not match with "AppliedOrAcademic" in "teacher questionnaire" but ClassID was the same [1] | 1,473 | 0.98% |
| G9_2010_ISD_SQd.csv | 154,569 | Duplicate StudentID [1] | 1,766 | 1.14% |
| G9_2010_TQ.csv | 6,373 | Variables "ClassID" and "AppliedOrAcademic" was different from any of the student's combination of "ClassID" and "Program" [2] | 193 | 1.51% |

Our research group managed to have a large statistical model of data after merging the given files. The individual merged files after each of the three steps contained all the related data about student questionnaire, student item responses and/or teacher questionnaire required to answer our research questions. The unrelated data, which represented 3%, 20% and 29% respectively at the first, second and third step of the original student questionnaire file, were dropped by writing relevant queries. The merged data files, number of records in the data files, criteria for the particular merge, number of records in the output file, and the percentage of dropped records in the original student questionnaire file are summarized in Table 2.

Table 2. Percentage of dropped records when data files were merged

| Merged files | Number of records in the data files | Criteria for the merge | Number of records in the output file | % of dropped records in the original student questionnaire file |
|---|---|---|---|---|
| Step 1 | | | | |
| G9_2010_ItemResponses.csv (I) | 150,186 | File I & II on the basis of RecID | 150,186 | 3 |
| G9_2010_ISD_SQd.csv (II) | 154,569 | | | |
| Step 2 | | | | |
| G9_2010_ItemResponses.csv (I) | 150,186 | File I & II on the basis of RecID and file III on the basis of ClassID | 132,772 | 20 |
| G9_2010_ISD_SQd.csv (II) | 154,569 | | | |
| G9_2010_TQ.csv (III) | 6,373 | | | |
| Step 3 | | | | |
| G9_2010_ItemResponses.csv (I) | 150,186 | File I & II on the basis of RecID and file III on the basis of ClassID and Program/ AppliedOrAcademic | 106,124 | 29 |
| G9_2010_ISD_SQd.csv (II) | 154,569 | | | |
| G9_2010_TQ.csv (III) | 6,373 | | | |

## 6. Discussion

The efforts involved in the merging of data file sources were successful. Initially, we thought that the data files were hard to analyze collectively because of their scattered nature, but we were able to achieve reasonable consistency through our merging model. Thus, a researcher working with data contained in multiple files has the opportunity to adopt variations of this model. The researcher has to be careful when dealing with multiple data files. The mechanism of data collection may differ in various available data files depending on the data collection instruments and their varying levels of accuracy affect the quality of merged data files (Voronin, 2006). Based on the experience of data merging in this project, we recommend data merging and the verification of a merged data file sources using SQL queries. Such verification of merged file sources provides an answer to where and why certain records are being dropped in the output file.

The problem of merging is obviously linked with the data integrity. When merging data file sources, the accuracy of each specific item in a data file becomes important. Large amounts of data typically have numerous data entry errors and duplicate entries about a variable, and it is difficult to identify which item in a data set contains faulty entry. These errors or duplicate entries affect the reliability and results of a study (Levitt, Aeppli, Potish, Lee, & Nierengarten, 1993). Certain models identify equivalent items using a complex, domain-dependent process (Hernández & Stolfo, 1998); however, they are difficult to implement in large-scale assessment research because data files may lack characteristic domain properties. With large-scale studies becoming commonplace in education, data merging and checking the integrity of the data is gaining importance. When merging different data file sources, especially when data is entered manually, data conflicts are likely to occur (Naumann & Häussle, 2002), and accuracy of each item of information is crucial. The typical method of data entry error correction in which a data entry operator just deletes or corrects an individual item is not applicable after data file sources have been merged. The verification and solution of these conflicts are important in the process of data management.

Several researchers have reported error rates in the single data entry method. For example, Prud'homme, Canner and Cutler (1989) stated that the keying error rate was 0.52% in a hypertension prevention trial. In another study, Spilker (1991) reported that the keying error rates resulted in 5% in a single data entry. Without an accurate identification and elimination of these data entry errors or duplicate records, descriptive statistics and various other aggregations will produce false or untrustworthy results. One possible and most commonly used solution could be double data entry to ensure that only correct data is entered into the database (McFadden, 1998). In the double data entry method, three steps are followed after the initial data entry: reentry, error detection, and error correction (Blumenstein, 1993). In the reentry step, another data entry operator enters the same data. In the error detection step, an automated process compares the two data sets and generates a list of discrepancies. In the last step, the list of discrepancies is checked and corrections are made accordingly. The double data entry method

requires a lot more resources. With a low data entry error rate, there is no need for double data entry in the database, and performing random checks on one out of twenty entries works well (Mullooly, 1990). The more we pursue the quality of data, the more it costs in terms of resources and time.

**References**


ACM (Association of Computing Machinery). (1991). SIGMOD (Special Interest Group on Management Of Data) record.

Allison, P. D. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Basilvesky, A., Sabourin, D., Hum, D., & Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Communications in Statistics, 14,* 371-394. http://dx.doi.org/10.1080/03610918508812445

Bitton, D., & DeWitt, D. J. (1983). Duplicate record elimination in large data files. *ACM Transactions on Database Systems, 8*(2), 255-265. http://dx.doi.org/10.1145/319983.319987

Blumenstein, B. A. (1993). Verifying keyed medical research data. *Statistics in Medicine, 12,* 1535-1542. http://dx.doi.org/10.1002/sim.4780121702

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American Journal of Education, 112*(4), 469-495. http://dx.doi.org/10.1086/505056

Coyer, S. M., & Gallo, A. M. (2005). Secondary analysis of data. *Journal of Pediatric Health Care, 19,* 60-63.

Crocker, R. K. (2002). *Learning outcomes: A critical review of the state of the field in Canada.* Ottawa: Canadian Educational Statistics Council.

Fellegi, I., & Sunter, A. A. (1969). Theory for record linkage. *American Statistical Association Journal, 64,* 1183-1210. http://dx.doi.org/10.1080/01621459.1969.10501049

Gorard, S. (2002). The role of secondary data in combining methodological approaches. *Educational Review, 54*(3), 231-237. http://dx.doi.org/10.1080/0013191022000016293

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery, 2*(1), 9-37. http://dx.doi.org/10.1023/A:1009761603038

Herron, D. G. (1989). Secondary data analysis: Research method for the clinical specialist. *Clinical Nurse Specialist, 3*(2), 66-69.

Hofferth, S. L. (2005). Secondary data analysis in family research. *Journal of Marriage and Family, 67,* 891-907. http://dx.doi.org/10.1111/j.1741-3737.2005.00182.x

Kornhaber, M. L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Educational Policy, 18*(1), 45-70. http://dx.doi.org/10.1177/0895904803260024

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21,* 23-30. http://dx.doi.org/10.1111/j.1745-3992.2002.tb00082.x

Levitt, S. H., Aeppli, D. M., Potish, R. A., Lee, C. K., & Nierengarten, M. E. (1993). Influences on inferences: Effect of errors in data on statistical evaluation. *Cancer, 72*(7), 2075-2082. http://dx.doi.org/10.1002/1097-0142(19931001)72:7<2075::AID-CNCR2820720704>3.0.CO;2-#

Louis, K. S. (1982). Sociologist as sleuth: Integrating methods in the RDU study. *The American Behavioral Scientist, 26*(1), 101-120. http://dx.doi.org/10.1177/000276482026001008

McFadden, E. (1998). *Data management in clinical trials.* New York: Elsevier.

Moriarty, H. J., Deatrick, J. A., Mahon, M. M., Feetham, S. L., Carroll, R. M., Shepard, M. P., & Orsi, A. J. (1999). Issues to consider when choosing and using large national databases for research of families. *Western Journal of Nursing Research, 21*(2), 143-153. http://dx.doi.org/10.1177/01939459922043794

Mullooly, J. P. (1990). The effects of data entry error: An analysis of partial verification. *Computers and Biomedical Research, 23,* 259-267. http://dx.doi.org/10.1016/0010-4809(90)90020-D

Nagy, P., Demeris, H., & van Barneveld, C. (2000). Priorities and values in accountability programs. *The Canadian Journal of Program Evaluation, 15,* 67-82.

Naumann, F., & Häussle, M. (2002). Declarative data merging with conflict resolution. In *Proceedings of ICIQ-02, Seventh International Conference on Information Quality* (pp. 212-224). Cambridge, MA: MIT.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556. http://dx.doi.org/10.3102/00346543074004525

Prud'homme, G. J., Canner, P. L., & Cutler, J. A. (1989). Quality assurance and monitoring in the hypertension prevention trial. *Controlled Clinical Trials, 10,* 84S-94S. http://dx.doi.org/10.1016/0197-2456(89)90044-5

Rew, L., Koniak-Griffin, D., Lewis, M., Miles, M., & O'Sullivan, A. (2000). Secondary data analysis: New perspective for adolescent research. *Nursing Outlook, 48,* 223-229. http://dx.doi.org/10.1067/mno.2000.104901

Roth, P. L., & Switzer, F. S. (1995). A Monte Carlo analysis of missing data techniques in a HRM setting. *Journal of Management, 21,* 1003-1023. http://dx.doi.org/10.1177/014920639502100511

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592. http://dx.doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association, 91,* 473-489. http://dx.doi.org/10.1080/01621459.1996.10476908

Simon, M., Roberts, N., Tierney, R., & Forgette-Giroux, R. (2007). Secondary analysis with minority group data: A research team's account of the challenges. *The Canadian Journal of Program Evaluation, 22*(3), 73-97.

Spilker, B. (1991). *Guide to clinical trials.* New York: Raven Press.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal, 338,* b2393. http://dx.doi.org/10.1136/bmj.b2393

Thakar, A. R., Szalay, A., Fekete, G., & Gray, J. (2008). The catalog archive server database management system. *Computing in Science & Engineering, 10*(1), 30-37. http://dx.doi.org/10.1109/MCSE.2008.15

van Barneveld, C., Stienstra, W., & Stewart, S. (2006). A content analysis of school board improvement plans in relation to the AIP model of educational accountability. *Canadian Journal of Education, 29*(3), 839-854. http://dx.doi.org/10.2307/20054198

Voronin, A. N. (2006). Synergetic methods of data merging in mathematical statistics. *Cybernetics and System Analysis, 42*(3), 320-327. http://dx.doi.org/10.1007/s10559-006-0068-5

Wang, Y. R., & Madnick, S. E. (1989). The inter-database instance identification problem in integrating autonomous systems. In *Proceedings of the Fifth International Conference on Data Engineering* (pp. 46-55). Los Angeles, California. http://dx.doi.org/10.1109/ICDE.1989.47199

Witta, E. L. (1992). *Seven methods of handling missing data using samples from a national database.* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA.

Wolfe, E. W., Ray, L. M., & Harris, D. C. (2004). A Rasch analysis of three measures of teacher perception generated from the school and staffing survey. *Educational and Psychological Measurement, 64*(5), 842-860. http://dx.doi.org/10.1177/0013164404263882